Cognitive Therapy and Research, Vol. 1, No. 2, 1977, pp. 161–175

Publication Prejudices: An Experimental Study of Confirmatory Bias in the Peer Review System

Michael J. Mahoney
Pennsylvania State University

Confirmatory bias is the tendency to emphasize and believe experiences which support one's views and to ignore or discredit those which do not. The effects of this tendency have been repeatedly documented in clinical research. However, its ramifications for the behavior of scientists have yet to be adequately explored. For example, although publication is a critical element in determining the contribution and impact of scientific findings, little research attention has been devoted to the variables operative in journal review policies. In the present study, 75 journal reviewers were asked to referee manuscripts which described identical experimental procedures but which reported positive, negative, mixed, or no results. In addition to showing poor interrater agreement, reviewers were strongly biased against manuscripts which reported results contrary to their theoretical perspective. The implications of these findings for epistemology and the peer review system are briefly addressed.

Cognitive psychologists have extensively documented the pervasiveness of error and distortion in human information processing (e.g., Neisser, 1967; Adams, 1967; Norman, 1969; Kintsch, 1970). Cognitively oriented clinical psychologists have also begun to note these fallibilities in a variety of dysfunctional patterns. Indeed, one of the major features of the more recent cognitive therapies has been the contention that many maladaptive behavior patterns are causally related to errors of thought and perception (e.g., Mahoney, 1974; Raimy, 1975; Beck, 1976). One particularly salient aspect of these erroneous cognitive processes might be termed *confirmatory bias.* This refers to the tendency for humans to seek out, attend to, and sometimes embellish experiences which support or "confirm" their beliefs. Confirmatory experiences are selectively welcomed and granted easy credibility. Disconfirmatory experiences, on the other hand, are often ignored, dis-

© 1977 Plenum Publishing Corp., 227 West 17th Street, New York, N.Y. 10011. To promote freer access to published material in the spirit of the 1976 Copyright Law, Plenum sells reprint articles from all its journals. This availability underlines the fact that no part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, microfilming, recording, or otherwise, without written permission of the publisher. Shipment is prompt; rate per article is $7.50.

credited, or treated with obvious defensiveness. A depressed client who thinks he is helpless may thus pay more attention to his failures and shortcomings. Instances of responsible control and success may be subjectively disregarded as unrepresentative or attributable to other forces. Similar examples could be offered for a wide range of clinical disorders (cf. Beck, 1976). The consequences of confirmatory bias are often tragic. By selectively "confirming" a maladaptive belief, the individual may lock himself into a vicious spiral of perception and performance. As the belief of helplessness gains support, for example, a client may initiate fewer attempts to control his own life—which leads to further opportunities for detecting helplessness and strengthening the belief.

The tragic effects of confirmatory bias are not, however, restricted to clinical disorders. In fact, as has been argued elsewhere (Mahoney, 1976), the most costly expression of this tendency may well be among scientists themselves. To the extent that researchers display this bias, our adequate understanding of the processes and parameters of human adaptation may be seriously jeopardized. If we selectively "find" or communicate only those data which support a given model of behavior, then our inquiry efforts will hardly be optimally effective. Despite the fact that confirmatory bias in scientists was first noted by Francis Bacon (1621/1960) over three centuries ago, precious little research has been devoted to the topic and the few extant studies have hardly challenged Bacon's observations. One study found that the vast majority of scientists drawn from a national sample showed a strong preference for "confirmatory" experiments (Mahoney & Kimper, 1976). Over half of these scientists did not even recognize disconfirmation (*modus tollens*) as a valid reasoning form! In another study the logical reasoning skills of 30 scientists were compared to those of 15 relatively uneducated Protestant ministers (Mahoney & DeMonbreun, 1977). Where there were performance differences, they tended to favor the ministers. Confirmatory bias was prevalent in both groups, but the ministers used disconfirmatory logic almost twice as often as the scientists.

The costs of this cognitive bias are perhaps nowhere as serious as in the area of scientific publication. The valuable contributions of a piece of research may be seriously threatened by a single act of human decision-making—namely, the judgment of a journal editor. There is substantial consensus among sociologists of science that the publication process is an integral part of contemporary science (Hagstrom, 1965; Ziman, 1968; Zuckerman & Merton, 1971; Cole & Cole, 1973). Unless his or her research is published, a scientist can have little hope of either personal advancement or recognized professional contribution. As documented in the research of Merton, Zuckerman, and others, journal publication has become the sine qua non of scientific achievement. In the absence of the public dissemina-

tion afforded by professional journals, a piece of research is often doomed to both obscurity and impotence in the growth of knowledge. Moreover, particularly in academic settings, lack of publication may seriously jeopardize the researcher's job security and continued research opportunities (Caplow & McGee, 1958; Dixon, 1973).

Given this integral function of publication, one might expect that the journal review process would have received extensive empirical scrutiny. In point of fact, the present article apparently represents the first controlled experimental exploration of this topic. This does not mean that publication policies have not been criticized; editorials and special articles have often cited the deficiencies of current review practices. Numerous allegations have been made about biases encountered in peer review. Moreover, a series of valuable post hoc analyses of editorial records have suggested that such biases may indeed be operative (Zuckerman & Merton, 1971; Merton, 1968; Zuckerman, 1970). For example, variables such as the author's prestige and institutional affiliation may significantly influence a reviewer's recommendation. Unfortunately, these correlational studies are limited in their implications and have been unable to investigate the relative influence of various factors within the research article itself.

In most research publications, four different components can usually be distinguished: an introduction, a description of experimental methodology, a summary of results, and an interpretation or discussion of the data. In journals which employ blind reviewing (wherein referees remain unaware of authorship and institutional affiliation), the above four components—supplemented by an abstract and a bibliography—may constitute the sole basis for reviewers' recommendations. To what extent do editors and referees weigh these various components in their evaluation? From an epistemological viewpoint, one might hope that the first two would far outrank the latter (Lakatos & Musgrave, 1970; Popper, 1972). That is, given that the researched question is relevant and the experimental methodology adequate, the obtained results—whatever they might be—should be of interest to the scientific community. Assuming that they are clearly and comprehensively described, the data should not be viewed prejudicially on the basis of whether they conform to current theoretical predictions. In fact, given that the logic of science should be more properly *falsificational* rather than *confirmational,* negative (or contratheoretical) results yield much more information than positive results (Weimer, 1977). It is only *unsuccessful* predictions which carry conclusive logical implications (Mahoney, 1976). Moreover, while they may disagree with the interpretation, the reviewers should not allow a discussion section to unduly bias their recommendations. They may, of course, exert their editorial prerogative in urging the author to publicly recognize (if not adopt) alternate data interpretations.

Although the ideal publication review system might emphasize relevance and methodological adequacy over data outcome and interpretation, this does not mean that such factors as writing style and suggested conclusions should have no bearing whatsoever on the editorial decision. However, these factors should exert less influence than the two primary empirical components. To what extent do referees adopt this philosophy in contemporary journal reviewing? This question was addressed in a study which asked referees to evaluate various experimentally manipulated manuscripts.

METHOD

Two basic factors were examined—the content of the reported data and their subsequent interpretation. Five groups of referees read manuscripts in which these variables were systematically altered. Introduction sections and methodologies were identical across articles. In two of the groups, however, the data reported were either consistent or inconsistent with the reviewer's presumed theoretical perspective. These opposite sets of data might be termed *positive* and *negative* results, respectively. A third group of reviewers was asked to evaluate the manuscript on the basis of its relevance and methodology alone—no results or data interpretation were offered. Two final groups of reviewers received manuscripts which contained relatively ambiguous or "mixed" results. In one group, these data were interpreted as being supportive of the reviewer's perspective; in the second, they were interpreted as contradictory. The five experimental groups, then, were as follows:

Group 1	Positive Results	No Discussion
Group 2	Negative Results	No Discussion
Group 3	No Results	No Discussion
Group 4	Mixed Results	Positive Discussion
Group 5	Mixed Results	Negative Discussion

The perspective of reviewers was inferred from their association with a journal which has been very energetic in advocating the refinement and expansion of applied behavioristic psychology—the *Journal of Applied Behavior Analysis.* Seventy-five referees were selected from the journal's list of guest reviewers for 1974. After random assignment to groups, they were invited to referee a brief research article purportedly submitted for publication in a compendium volume on "Current Issues in Behavior Modification." In those groups where manuscript components were absent, referees were told that these parts were in preparation and, due to a tight deadline, would be evaluated separately by the editor when they were received. Referees given partial manuscripts were asked to evaluate the merits of the

article in its inchoate state. All reviewers were asked to use the evaluation criteria explicitly outlined by the *Journal of Applied Behavior Analysis.*

The manuscript employed was a brief report of a study examining the effects of extrinsic reinforcement on intrinsic interest. This topic has recently become very controversial in psychology (Deci, 1971, 1972; Lepper, Greene, & Nisbett, 1973; Levine & Fasnacht, 1974). A group of psychologists has argued that the popular behavioristic strategy of reinforcement may have serious negative side effects. According to their argument, rewarding an individual for some behavior may sometimes cause him to devalue its intrinsic merits and thereby undermine his interest in it after reinforcement incentives have been terminated. Thus, the currently popular practice of rewarding a child for classroom performances may lead to a devaluation of academic tasks. Behaviorists have responded to these allegations with energetic and almost uniform denial, arguing that reward often enhances rather than undermines intrinsic interest.

The experimental manuscript described a fictitious experiment addressing this issue. After noting the timely relevance of the question, a methodology section described procedures aimed at evaluating the hypothesis. Three groups of preschool children were purportedly studied. Data from their performance of two play session activities (pressed wood puzzles and children's books) were said to have been collected and scored via closed-circuit television and independent raters. After completing a 2-week baseline to evaluate initial performance (interest) rates, the three groups were alleged to have experienced different experimental procedures. During the next 4 weeks, one group was said to have been rewarded with toy prizes for increments in their puzzle-solving behavior. The second group served as an exposure control and simply continued to have access to the two play activities. A third group was the formal control condition for both reinforcement and exposure; these children did not have access to the activities for the 4-week interval. To better evaluate any enduring changes in performance, a 6-week hiatus ensued followed by a 4-week follow-up asssessment in which all three groups were allegedly given daily opportunities to engage in either activity (without any further reward).

Each manuscript contained an identical bibliography in which half the references were either supportive or critical of behavior modification. Referees were blind to authorship and institution. To ensure that the introduction and methodology were within the bounds of acceptability for the reference journal, the approval of one of its associate editors was obtained. Manuscripts for the first two groups (positive and negative results) presented individual and group data for the three experimental conditions. In addition, Figure 1 was presented. Since the hypothesis at issue dealt with performance *after* the termination of reward, the critical data were those of the third (follow-up) experimental phase. For positive results referees, curve

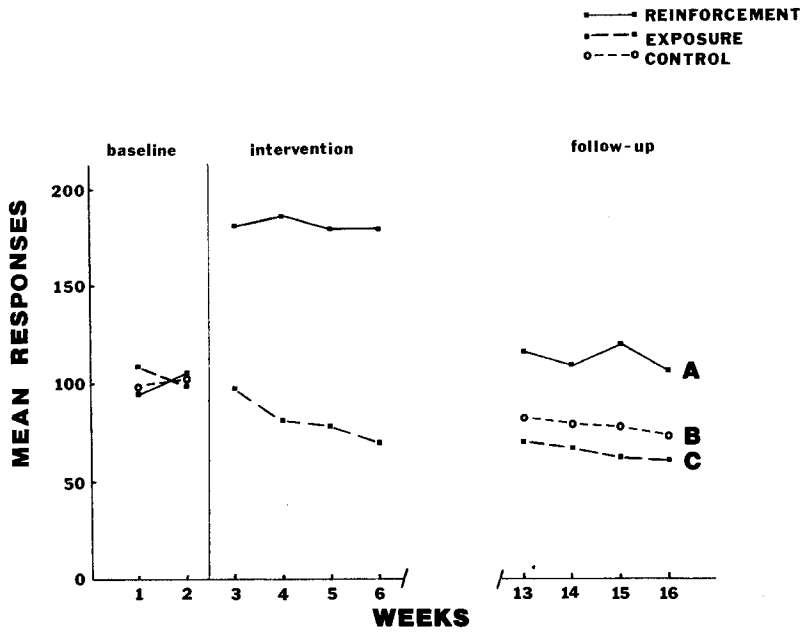

Fig. 1. Figure sent with positive results manuscripts. For negative results manuscripts, the labels A and C were inverted at follow-up.

A was labeled "Reinforcement," curve B was labeled "Control," and curve C was labeled "Exposure." Negative results referees received the same tables and figures except that the labels and data for curves A and C were reversed.

Referees in the final two groups (4 and 5) received manuscripts depicting results which were substantially more ambiguous than the foregoing. Although curve B remained unchanged, curves A and C reflected data which were equivocal to the hypothesis. As shown in Figure 2, the absolute altitude of A was greater than that of C, but its slope suggested a trend of performance decrement. Thus, depending on whether one emphasized altitude or slope, a different conclusion might be drawn. This engineered ambiguity facilitated the construction of two opposite discussions—one claiming that reward did not undermine performance (group 4) and one arguing that it did (group 5). Since it was impossible to predict referees' perceptions of ambiguous data, the curve labels were reversed for half the manuscripts within each group. Thus, 6 of the 13 reviewers in group 4 and 7 of the 14 in group 5 received tables and a figure in which curve A represented the reinforcement condition; for the remainder, A depicted the exposure group.

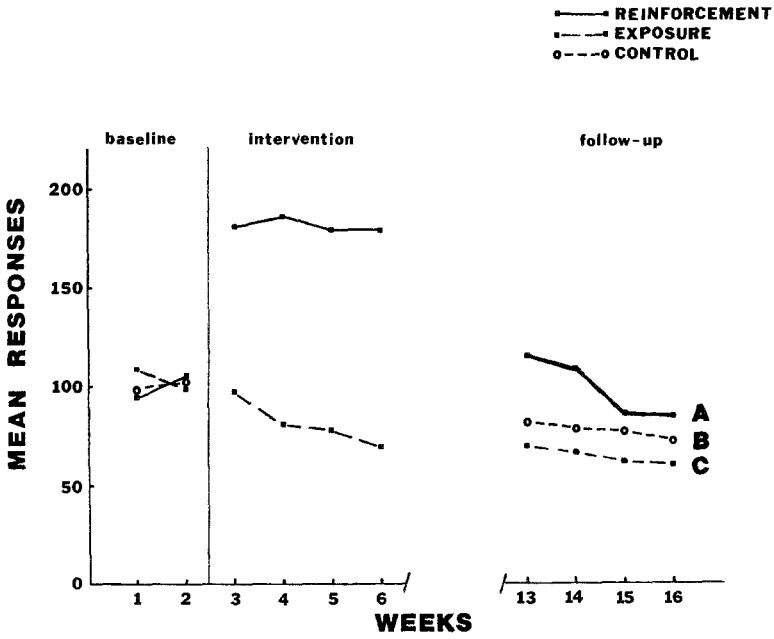

Fig. 2. Figure sent with mixed results manuscripts.

In addition to an open-ended referee form, reviewers were asked to rate the manuscript on five factors: topic relevance, methodology, data presentation, discussion, and overall scientific contribution. A 4-point rating scale was used (poor, marginal, adequate, good). Purportedly to aid editorial decision, reviewers were also asked to quantify their summary recommendation on another 4-point scale (accept, accept with minor revisions, accept with major revisions, reject). All referees were apprised of a tight deadline and asked to return their evaluations within 45 days.

RESULTS

Out of the initial sample of 75 (15 per group), 67 reviews were obtained (see Table II for group *N*s). Forty-six percent of these were received by deadline. Mixed results manuscripts (groups 4 and 5) were returned earlier than those containing clear-cut results (groups 1 and 2). However, since the mixed-results manuscripts also contained a discussion section, it was not possible to identify the source of this variance. The summary recommendations of punctual reviews were marginally more negative than those given in tardy evaluation ($t = 1.81$, $p < .08$, two-tailed).

Table I. Group Data and Statistical Analyses

Variable		1 Positive results	2 Negative results	3 No results	4 Mixed, positive discussion	5 Mixed, negative discussion
Percent returned by deadline		25.0	36.0	43.0	46.0	71.0
Clear-cut (1 + 2) versus mixed results (4 + 5) Comparison $X^2 = 7{,}0, p < .01$						
Topic relevance rating	Mean	5.2	4.9	5.1	5.2	5.3
	SD	1.0	2.0	1.5	1.3	1.7
F (4,58) = .14 Contrast 1 (1 vs. 2), t = .53; Contrast 2 (1 vs. 4 + 5), t = .05; Contrast 3 (2 vs. 3), t = .34; Contrast 4 (4 vs. 5), t = .19						
Methodology rating	Mean	4.2	2.4	3.4	2.5	2.7
	SD	1.9	2.4	2.2	1.5	2.4
F (4,58) = 1.35 C-1 $t = 2.06, p < .05$, 95% confidence intervals = .34 to 3.78; C-2 t = 2.16, $p < .05$, c.i. = .68 to 3.64; C-3 t = 1.10; C-4 t = .24						
Data presentation rating	Mean	4.3	2.6	–	1.3	2.0
	SD	.9	2.0	–	1.4	1.9
F (3,46) = 6.44, $p < .01$ C-1 $t = 2.58, p < .02$, c.i. = 1.23 to 3.93; C-2 $t = 4.61, p < .001$, c.i. = 3.44 to 5.78; C-4 t = 1.12						
Discussion rating	Mean	–	–	–	1.3	.9
	SD	–	–	–	1.3	1.3
Scientific contribution rating	Mean	4.3	2.4	4.5	1.6	1.7
	SD	1.4	2.2	2.4	1.3	2.1
F (4,51) = 5.78, $p < .01$ C-1 $t = 2.35, p < .03$, c.i. = .66 to 4.04; C-2 $t = 3.76, p < .001$, c.i. = 2.31 to 5.21; C-3 $t = 2.55, p < .02$, c.i. = .86 to 4.24; C-4 t = .11						
Summary recommendation	Mean	3.2	1.8	3.4	.5	1.4
	SD	1.4	1.9	2.3	.9	1.7
F (4,61) = 6.69, $p < .01$ C-1 $t = 2.21, p < .05$, c.i. = .87 to 3.55; C-2 $t = 3.91, p < .001$, c.i. = 2.75 to 5.07; C-3 $t = 2.48; p < .02$, c.i. = 1.14 to 3.82; C-4 t = 1.33						

Referee evaluations and summary recommendations were scored on a Likert scale with 0 as the lowest rating, 2 and 4 as intermediate, and 6 as the highest. Reviewer responses were subjected to analyses of variance with three planned orthogonal contrasts: positive versus negative data (group 1 vs. 2), positive discussion versus negative discussion (4 vs. 5), and no results versus results (3 vs. 1 + 2 + 4 + 5). The chosen alpha level for statistical significance was .05, two-tailed, and the five factors were analyzed separately. A summary of the data and these analyses is presented in Table I. Individual referee responses are given in Table II. Analyses of intergroup homogenity of variance indicated that this assumption was warranted on all factors except final recommendation. For this factor, nonparametric statistics were employed (see Table I).

Referee ratings of the first factor, Topic Relevance, did not differ across groups. Similarly, planned orthogonal comparisons between specific groups failed to reveal significant differences. The average rating across

Table II. Individual Referee Ratings by Group

	Group 1 Positive results					Group 2 Negative results					Group 3 No results				Group 4 Mixed results, positive discussion						Group 5 Mixed results, negative discussion					
Reviewer	Relevance	Methodology	Data	Contribution	Recommendation	Relevance	Methodology	Data	Contribution	Recommendation	Relevance	Methodology	Contribution	Recommendation	Relevance	Methodology	Data	Discussion	Contribution	Recommendation	Relevance	Methodology	Data	Discussion	Contribution	Recommendation
1	6	6	6	6	4	6	0	0	0	0	6	6	6	4	6	2	2	2	2	0	6	4	2	0	4	2
2	6	3	4	3	2	6	4	2	–	4	2	4	2	0	6	4	2	2	4	2	6	6	6	4	6	6
3	6	6	4	6	4	4	2	0	2	1	–	–	–	6	4	2	0	2	2	0	4	0	2	2	–	1
4	4	4	4	4	2	6	0	6	0	0	6	0	6	2	6	4	4	0	2	2	6	6	4	0	0	1
5	6	0	4	2	0	6	0	2	2	0	6	0	0	0	6	2	2	0	2	2	4	0	0	0	0	0
6	4	4	4	4	4	2	0	0	0	0	2	2	–	2	6	4	0	2	0	0	6	0	2	0	0	0
7	4	6	4	4	4	2	0	2	2	0	6	4	6	4	–	–	–	–	–	0	6	4	0	0	0	2
8	4	4	4	4	4	6	4	4	4	4	6	4	6	6	2	4	2	4	2	0	6	0	4	0	4	1
9	6	3	3	–	2	6	4	4	4	4	4	4	4	4	6	2	0	0	0	0	6	4	2	2	–	4
10	–	–	–	–	5	6	4	4	–	–	6	4	6	4	6	3	3	2	3	1	6	2	2	0	0	0
11	6	6	6	6	4	6	0	0	0	0	6	6	6	6	4	0	0	0	0	0	6	6	0	0	2	0
12	–	–	–	–	4	0	4	4	–	2	6	6	6	6	4	0	0	0	0	0	0	0	0	0	0	0
13						6	6	4	6	4	6	4	6	4	6	2	0	2	2	0	6	4	4	2	2	2
14						6	6	4	6	4	4	0	0	0							6	2	0	2	2	1

subjects was a high 5.11. In their evaluations of Methodology, however, group differences began to emerge. With identical experimental procedures, a positive results manuscript was rated as methodologically better than one reporting negative results. The difference between manuscripts with and without results sections did not attain statistical significance.

Scores on the third factor, Data Presentation, were also affected by the direction of the reported results. Data sections which offered evidence supportive of behavior modification were rated as significantly better than those which were critical. The nature of the discussion section did not appear to influence referees' evaluations of ambiguous findings.

Estimates of overall Scientific Contribution reflected a parallel trend. Positive results manuscripts were rated as being much more contributory than negative results papers. In the absence of any reported results, reviewers rated a manuscript much higher than when it reported its findings. Again, the discussion section did not appear to be influential.[1]

[1]Because several reviewers failed to rate the manuscript on this dimension, its N and degrees of freedom were slightly reduced (see Table II).

The critical dependent variable, of course, was that which is most relevant to publication—namely, the reviewer's Summary Recommendation. Here again a familiar pattern emerged. Identical manuscripts suffered very different fates depending on the direction of their data. When they were positive, the usual recommendation was to accept with moderate revisions. Negative results earned a significantly lower evaluation, with the average reviewer urging either rejection or major revision. Referees who were not given any results were much more generous in their recommendations than reviewers who read a results section. Mixed results manuscripts were consistently rejected without any apparent influence by their manner of interpretation.

An unplanned dependent variable was made possible by an overlooked typographical error in the experimental manuscript. Although the method section stated that subjects had been randomly divided into triads, it mistakenly contended that a total of eight subjects were employed. Since eight is not evenly divisible by three, the alleged procedure was impossible. Moreover, referees who received a results section viewed a table depicting data for 12 subjects (which was the intended sample size). Reviewers in group 3 (No Results) thus saw one instance of contradiction and all other reviewers saw two. Analyses were performed to evaluate whether the five groups noted this contradiction with equal frequency. Of the individuals who read a positive results manuscript (group 1), only 25% noted the above problem. When reading negative results (group 2), however, 71.4% of the reviewers detected the contradiction. By Fisher's Exact Probability Test, this difference must be considered substantial ($p < .05$). Reviewers who did not see any results (and therefore had only one index of contradiction) still noted the discrepancy 35.7% of the time. In the mixed results groups (4 and 5), 53.8% and 28.6% of the reviewers called attention to the contradiction.

Although direction of influence cannot be identified, correlations among the major dependent variables revealed some interesting relationships. Referees who read positive or negative results manuscripts showed a marked covariation between their methodology ratings and summary recommendations ($r = .94$). Evaluations of the data section were also positively correlated with recommendations ($r = .56$) and, most interestingly, referees' data ratings were apparently related to their methodology ratings ($r = .60$). Taken together, these correlations suggest the possibility of a "halo effect" in which manuscript components share a common valence rather than being rated independently.

Referees showed relatively modest agreement with one another in their component ratings and summary recommendations. On topic relevance, average interrater agreement across all five groups was a meager $-.07$ (intraclass correlation coefficient (a)). Likewise, their evaluations of methodology showed little consensus ($a = .03$). Ratings of data presenta-

tion, scientific contribution, and summary recommendation showed somewhat higher interreferee agreement, but were again modest (all 3 $a = .30$). In the last two groups, evaluations of the discussion section showed little consensus ($a = .01$). A postexperimental questionnaire asked referees to predict their degree of reliability with other reviewers on the various items. Their average predictions—contrasted with the obtained values—were as follows:

Factor	Self-Prediction	Obtained Value
Relevance	.74	−.07
Methodology	.69	.03
Data Presentation (1,2,4,5)	.72	.30
Discussion (4,5)	.72	−.01
Scientific Contribution	.72	.30
Summary Recommendation	.72	.30

The difference between self-predicted and obtained reliabilities is striking.

In addition to their standard rating form, referees were invited to submit comments and suggestions for the author. It should be noted that many of the reviewers spent considerable time and effort in executing their task. Several sent relevant bibliographies for the hypothetical author's use, two forwarded related reprints and thesis abstracts, and one referee submitted a hand-calculated analysis of variance on the fabricated data. Moreover, their reviews were frequently constructive—even when critical—and most reflected considerable examination. They were often several pages in length, with one almost 2,000 words long.

Representative referee comments are presented in Table III. Several patterns are apparent. First, wide variability was again encountered both within and between groups. Looking only at the comments, one would hardly think that very similar or even identical manuscripts were being evaluated. A second pattern was the frequent feelings of awkwardness reported by referees who received incomplete manuscripts. They often complained about the handicap of missing components and qualified their remarks by such phrases as "assuming that the Discussion is reasonable." Finally, the emphasis placed on data content is again reflected in referee comments. Ambiguous data are explicitly devalued as lacking scientific contribution.

After the reviews had been returned, referees were sent a letter informing them of the nature of the project and asking them to fill out a brief questionnaire. The latter asked them to rate the importance of research on the journal review process, to predict their reliability with other referees, and to describe their prior suspicions and subsequent emotional reactions regarding the experimental nature of the project. Of the 57 individuals (85%) who responded, 4 had been somewhat suspicious and 13 expressed

Table III. Representative Referee Comments

Group	Referee comment
(1) Positive results	"A very fine study. . . . I have not seen the Discussion section but I don't see how it could be very far off the mark." "An excellent paper . . . it definitely merits publishing. I find little to criticize. The topic is excellent and very relevant, the design is quite adequate, and the style is very good." "It's a bit difficult to review this sort of study without the discussion section!"
(2) Negative results	"There are so many problems with this paper, it is difficult to decide where to begin. While I have not seen the discussion section, I can't think of what would be there to save this paper." "The paper [is] perpetrating a serious, mistaken conclusion by unwary readers." "I would hope that the authors avoid making . . . wild overgeneralizations." "Accept as exploratory study if [the] discussion includes alternate explanations of the data."
(3) No results	"Very good. Well done. If the Results and Discussion . . . are as well written . . . I definitely recommend publication. "I would suggest that the only . . . results which would merit publication would be if the performance of reinforcement *S*s deteriorates." "I felt rather strange reviewing this article in its incomplete form." "I have had very mixed emotions (mostly 'displeases') about reviewing such an incomplete manuscript. Personally, I don't see how anyone can write the Introduction and Method without first having the Results." "Reading half of a journal article . . . must be analogous with the situation proposed in, 'What is worse than biting into an apple and finding you ate a whole worm?' 'Biting into an apple and finding you ate half a worm.' "
(4) Mixed results, positive discussion	"There is sufficient ambiguity in the data so that any conclusions . . . could not be made with any degree of certainty." "This study presupposes that the 'undermining' hypothesis warrants an involved empirical evaluation. Disproving insubstantial theoretical hypotheses is generally not considered an adequate rationale for publication." "The author's conclusions are at best inconclusive. . . . I do not advise acceptance of the article."
(5) Mixed results, negative discussion	"This is a seriously flawed study, both in conceptualization and analysis." "Either I have missed something or this is a bizarre article. . . . Reject." "I find no fault in the method or data analysis. . . . My reservations, then, have to do with the introduction and discussion." "This report is a classic example of hypothesis myopia. . . . The authors have drawn conclusions which are completely unsupported by their own findings."

negative reactions, primarily regarding the deception involved. The vast majority reported a mixture of surprise, curiosity, and commendation. Ratings of the importance of this type of research were consistently high and, without exception, referees asked to see a copy of the study's results.

DISCUSSION

Two general conclusions may be drawn from the present study. Within the constraints of its subject population and methodology, it was found that (a) referee evaluations may be dramatically influenced by such factors as experimental outcome, and (b) interreferee agreement may be extremely low on factors relating to manuscript evaluation. What are the implications of these findings? The answer to that question is neither simple nor straightforward. First, how should we deal with the apparent prejudice against "negative" or disconfirming results? I have argued elsewhere that this bias may be one of the most pernicious and counterproductive elements in the social sciences (Mahoney, 1976). One possible solution might be to ask referees to evaluate the relevance and methodology of an experiment without seeing either its results or their interpretation. While this might be a dramatic improvement, it raises other evaluative problems. How does one deal with the fact that referees may show very little agreement on these topics? Training them might produce better consensus, but consensus is not necessarily unprejudiced. Referees might achieve perfect agreement by simply sharing the same ideological or methodological biases.

The American Psychological Association (1966) recommends that psychological tests "should report evidence of reliability that permits the reader to judge whether scores are sufficiently dependable" (p. 27). Various indices of validity are also requested. What has been apparently overlooked is the fact that *peeer review is a form of evaluative testing.* Journal editors and referees are asked to judge a manuscript in terms of its scientific "worthiness." Unfortunately, the criteria for scientific worth are hardly unequivocal and appear to be currently undergoing a drastic reappraisal (Weimer, 1977; Mahoney, 1976). More embarrassing, perhaps, is the realization that we have developed elaborate standards for evaluating various psychological instruments and yet have exempted the most pervasive and critical instrument in science—i.e., the scientist. Have we presumed that it is "naturally" reliable and objective? With our vast literatures on information processing and social psychology, have we assumed that scientists are somehow unaffected by the processes which appear to be so common in other members of the species?

Confirmatory bias is not, of course, the only potential source of prejudice in peer review. A recent experimental study has, for example, shown that citing your own "in press" publications may significantly enhance your chances of earning a reviewer's approval (Mahoney, Kazdin, & Kenigsberg, 1975). The ironic feature of confirmatory bias is the fact that it is fundamentally illogical. Positive results and negative results experiments are not equivalent in their logical implications. In fact, while they have unquestionable bearing on the subjective aspects of belief, *successful experiments have no necessary logical bearing on the truth status of their source* (i.e., a theory or hypothesis). As counterintuitive as this may seem, it is a clear consequence of logical analysis (cf. Popper, 1972; Weimer, 1977; Mahoney, 1976). *It is only negative results (contrary-to-prediction) experiments which carry logical implications.* The reasons for this are simple and are outlined in the above-mentioned sources. Despite this clear mandate from logic, however, our research programs and publications policies continue in their dogmatically confirmatory tradition. They offer ample testimony to Bacon's (1621/1960) astute observation that "the human intellect... is more moved and excited by affirmatives than by negatives."

Without further scrutiny of the purposes and processes of peer review, we are left with little to defend it other than tradition. While the journal review process is only one aspect of contemporary science, it is probably one of the more critical. Ironically, it is also one of the most neglected. We have assumed that peer review is an adequate and objective process in its present form, and there has been little effort to challenge that assumption. The present article aspires to such a challenge. Its premise is simply that we do not adequately understand either the processes or the effects of our conventional practices, and—more importantly—that we are negligent if we fail to study them. Until we subject our publication policies to the same empirical scrutiny allotted other research topics, we have little means for assessing or refining this pivotal link in the chain of empirical knowledge.

REFERENCES

Adams, J. A. *Human memory.* New York: McGraw-Hill, 1967.

American Psychological Association. *Standards for educational and psychological tests and manuals.* Washington, D.C.: American Psychological Association, 1966.

Bacon, F. *Novum organum.* New York: Bobbs-Merrill, 1960. (Originally published, 1621.)

Beck, A. T. *Cognitive therapy and the emotional disorders.* New York: International Universities Press, 1976.

Caplow, T., & McGee, R. J. *The academic marketplace.* New York: Basic Books, 1958.

Cole, J. R., & Cole, S. *Social stratification in science.* Chicago: University of Chicago Press, 1973.

Deci, E. L. Effects of externally mediated rewards on intrinsic motivation. *Journal of Personality and Social Psychology,* 1971, *18,* 105-115.

Deci, E. L. Intrinsic motivation, extrinsic reinforcement, and inequity. *Journal of Personality and Social Psychology,* 1972, *22,* 113-130.

Dixon, B. *What is science for?* New York: Harper & Row, 1973.

Hagstrom, W. O. *The scientific community.* New York: Basic Books, 1965.

Kintsch, W. *Learning, memory, and conceptual processes.* New York: Wiley, 1970.

Lakatos, I., & Musgrave, A. (Eds.), *Criticism and the growth of knowledge.* London: Cambridge University Press, 1970.

Lepper, M. R., Greene, D., & Nisbett, R. E. Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. *Journal of Personality and Social Psychology,* 1973, *28,* 129-137.

Levine, F. M., & Fasnacht, G. Token rewards may lead to token learning. *American Psychologist,* 1974, *29,* 814-820.

Mahoney, M. J. *Cognition and behavior modification.* Cambridge, Massachusetts: Ballinger, 1974.

Mahoney, M. J. *Scientist as subject: The psychological imperative.* Cambridge, Massachusetts: Ballinger, 1976.

Mahoney, M. J., & DeMonbreun, B. G. Confirmatory bias in scientists and non-scientists. *Cognitive Therapy and Research,* 1977 (in press).

Mahoney, M. J., Kazdin, A. E., & Kenigsberg, M. *Getting published: The effects of self-citation and institutional affiliation.* Unpublished manuscript, Pennsylvania State University, 1975.

Mahoney, M. J., & Kimper, T. P. From ethics to logic: A survey of scientists. In M. J. Mahoney, *Scientist as subject.* Cambridge, Massachusetts: Ballinger, 1976. Pp. 187-193.

Merton, R. K. *Social theory and social structure.* New York: Free Press, 1968.

Neisser, V. *Cognitive psychology.* New York: Appleton-Century-Crofts, 1967.

Norman, D. A. *Memory and attention.* New York: Wiley, 1969.

Popper, K. R. *Objective knowledge: An evaluative approach.* London: Oxford University Press, 1972.

Raimy, V. *Misunderstandings of the self.* San Francisco: Jossey-Bass, 1975.

Weimer, W. B. *Psychology and the conceptual foundations of science.* Hillsdale, New Jersey; Lawrence Erlbaum Associates, 1977.

Ziman, J. *Public knowledge: The social dimension of science.* London: Cambridge University Press, 1968.

Zuckerman, H. Stratification in American science. *Sociological Inquiry,* 1970, *40,* 235-257.

Zuckerman, H., & Merton, R. K. Patterns of evaluation in science: Institutionalization, structure and functions of the referee system. *Minerva,* 1971, *9,* 66-100.